Nonlinear Audio Systems Identification Through Audio Input Gaussianization

Imen Mezghani-Marrakchi, Gaël Mahé, Sonia Djaziri-Larbi, Mériem Jaïdane, Monia Turki-Hadj Alouane

Nonlinear audio systems identification through audio input Gaussianization

Imen Mezghani-Marrakchi, Gaël Mahé, Sonia Djaziri-Larbi, Mériem Jaïdane, Monia Turki-Hadj Alouane

Abstract

Nonlinear audio system identification generally relies on Gaussianity, whiteness and stationarity hypothesis on the input signal, although audio signals are non-Gaussian, highly correlated and non-stationary. However, since the physical behavior of nonlinear audio systems is input-dependent, they should be identified using natural audio signals (speech or music) as input, instead of artificial signals (sweeps or noise) as usually done.

We propose an identification scheme that conditions audio signals to fit the desired properties for an efficient identification. The identification system consists in (1) a Gaussianization step that makes the signal near-Gaussian under a perceptual constraint; (2) a predictor filterbank that whitens the signal; (3) an orthonormalization step that enhances the statistical properties of the input vector of the last step, under a Gaussianity hypothesis; (4) an adaptive nonlinear model.

The proposed scheme enhances the convergence rate of the identification and reduces the steady state identification error, compared to other schemes, for example the classical adaptive nonlinear identification.

Imen Mezghani-Marrakchi, Sonia Djaziri-Larbi, Mériem Jaïdane and Monia Turki-Hadj Alouane are with Université Tunis El Manar, Ecole Nationale d'Ingénieurs de Tunis (ENIT), Signals and Systems Lab (U2S), BP 37, 1002 Tunis-Belvédère, Tunisia. e-mail: mezghani_imen@yahoo.fr, (sonia.larbi, m.turki)@enit.rnu.tn, meriem.jaidane@planet.tn

Gaël Mahé is with the Laboratory of Informatics Paris Descartes (LIPADE) in the University Paris Descartes, 45 rue des Saints Pères, 75270 Paris Cedex 06, France. e-mail: gael.mahe@mi.parisdescartes.fr

I. INTRODUCTION

Nonlinear behavior of acoustic systems is a problem encountered in various audio applications such as cellular phones, video conferencing systems and public address sound reinforcement. Low-cost audio equipments and constraints of portable communication systems accentuate this phenomenon. These distortions are a superposition of different mechanical, electrical and acoustical effects, which can be modeled by polynomial models for memoryless systems and by Volterra filters [1], [2], [3] for systems with memory.

For example, loudspeakers are modeled by Volterra filters with a nonlinearity order of 2 [4] to 3 [5], [6]. Audio amplifiers have also a nonlinear behavior, which was emulated in [7] for a tube preamp (as used by electric guitars) by a Volterra model with nonlinearity order 10.

Classical identification algorithms of nonlinear audio systems use synthetic signals as inputs. In [8], [9], and in a context of nonlinear acoustic echo cancellation, the nonlinear echo path has been identified with a stationary white Gaussian input. For loudspeakers, classical input signals for identification are multitones [10], sine sweeps, Maximum Length Sequences (MLS), wide MLS (interleaving zeros between ± 1) and multiple noises with modulus equal to 1 [11].

However, the physical behavior of nonlinear audio systems is input-dependent. This was stressed in [12] for speech communication systems: classical steady-state measurements (sweeps, tones, noises...) are not sufficient to predict the subjective performance of a system, so that they should be replaced by speech-like test stimuli. In a more physical approach, Klippel [13], [14] showed the relationship between the properties of the input signal and the physical behavior of a loudspeaker. For example, the voice coil heating, which generates nonlinear distortions, depends on the spectral properties of the stimulus. As a consequence, a full dynamic measurement, that excites all the nonlinearities to be measured, is performed with audio-like stimuli.

Hence, audio nonlinear systems should be identified when they are excited by their real inputs (natural audio signals). But the properties of audio signals make them unsuitable for classical identification algorithms, since they are generally non-Gaussian, non-stationary and highly correlated. This point was raised in [15] for the efficiency measurement of audio amplifiers: while synthetic signals cause a different system behavior than audio, the non-stationarity of audio signals makes them difficult to use as test input.

Several studies take into account some of the natural inputs properties. In [16], the authors proposed a decorrelation filter to turn white the input, which is useful to pilot the adaptive filter. However, the non-commutativity of the decorrelation filter, which is linear, and the nonlinear Volterra filter limits the validity

of this method. An identification method was proposed in [17] with high algorithmic complexity, for stationary, Gaussian but correlated inputs. It consists in a prediction step followed by an orthogonalization step. This method was tested in the adaptive identification case for a Volterra system of low order and in the particular case of an AR(1)[1] Gaussian process. An enhancement was achieved for both transient and steady states. Nevertheless, this method was not validated for high order systems nor for non-Gaussian and non-stationary inputs.

Thus, we propose to take fully into account the properties of audio signals, namely non-Gaussianity, non-stationarity and high correlation, in the identification of nonlinear audio systems. In section II, we point out the importance of Gaussianity in identification algorithms. Then, we propose in section III a "Gaussianization" algorithm that aims at making an audio signal more Gaussian without changing its perceptual properties. In section IV, we present a new identification structure based on Gaussianity and taking into account the correlation and the non stationarity of audio signals. Finally, in section IV, we present a simulation study and discuss the simulation results.

II. Ill-conditioning in nonlinear system identification

Speech signals have been shown to be near-Laplacian, whereas the distribution of music signals depends on the type and the number of instruments, and tends to be Gaussian when several instruments are involved [18]. Audio signals can be considered as generalized Gaussian processes, which distribution varies from Gaussian to Laplacian.

The PDF[2] $p(x)$ of a generalized Gaussian process is given by

$$p(x) = \frac{\nu.\eta(\nu,\sigma)}{2.\Gamma(1/\nu)} \exp[-[\eta(\nu,\sigma).|x|]^{\nu}], \tag{1}$$

where σ^2 is the variance of x and

$$\eta(\nu,\sigma) = \frac{1}{\sigma}\left[\frac{\Gamma(3/\nu)}{\Gamma(1/\nu)}\right]^{1/2}, \tag{2}$$

where $\Gamma(.)$ is the Gamma function. The larger is the ν factor the flatter is the PDF. The PDF

- is Laplacian for $\nu = 1$;
- is Gaussian for $\nu = 2$;
- tends to an impulse function for $\nu \longrightarrow 0$;
- tends to a uniform distribution for $\nu \longrightarrow +\infty$.

[1]AR(n): autoregressive process with order n.

[2]PDF : probability density function.

A theoretical analysis is presented here to exhibit the importance of input Gaussianity in nonlinear systems identification. In the literature, nonlinear distortions are generally modeled by polynomial structures (for nonlinear memoryless systems) or truncated Volterra series (for nonlinear systems with memory) which are identified using optimal or adaptive algorithms. These algorithms are sensitive to the ill-conditioning of the observation matrix, even if the input is white. This concerns a matrix inversion problem for optimal identification and a convergence problem in the adaptive case.

In the following, a particular attention is paid to the influence of the input PDF on the conditioning of these observation matrices.

A. Polynomial systems

In the case of a polynomial system identification, the observation vector is $X_k = [1, x_k, x_k^2, ..., x_k^N]^\top$ where x_k is the input signal and N refers to the polynomial order. For a stationary process,

$$\mathbf{C_x} = E[X_k X_k^\top],$$

is the symmetric matrix defined by

$$\mathbf{C_x} = \begin{bmatrix} 1 & m_1 & m_2 & . & . & m_N \\ m_1 & m_2 & m_3 & . & . & m_{N+1} \\ m_2 & m_3 & . & . & m_{N+1} & m_{N+2} \\ . & . & . & m_{N+1} & m_{N+2} & . \\ . & . & m_{N+1} & m_{N+2} & . & . \\ m_N & m_{N+1} & m_{N+2} & . & . & m_{2N} \end{bmatrix}, \tag{3}$$

where $m_i = \mathrm{E}[x_k^i]$ is the i^{th} order moment and E[.] denotes the expectation value. With language misuse and for simplicity, we will call the matrix $\mathbf{C_x}$ "correlation matrix".

The identification performance are closely related to the conditioning of the matrix $\mathbf{C_x}$ which depends on the PDF of the input signal x_k. The conditioning of $\mathbf{C_x}$ is evaluated through its logarithmic condition number [19]

$$K(\mathbf{C_x}) = \log_{10}\left(\frac{|\lambda_{max}|}{|\lambda_{min}|}\right), \tag{4}$$

where λ_{max} and λ_{min} are respectively the largest and the smallest eigenvalues of the matrix $\mathbf{C_x}$.

We compared $K(\mathbf{C_x})$ for various orders and values of the form factor ν of a generalized Gaussian PDF (Fig. 1). The theoretical values were computed for the Gaussian, Laplacian and uniform cases.

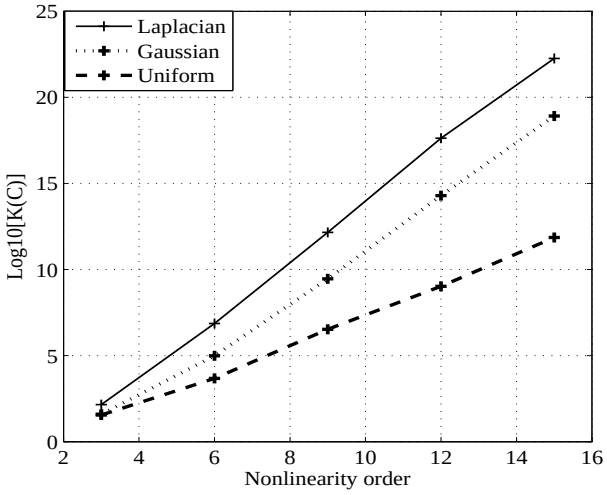

Fig. 1. For a polynomial system, condition number $K(\mathbf{C_x})$ according to the nonlinearity order, for different generalized Gaussian processes.

Applying the Price theorem [20] on a zero mean Gaussian process x^g, we may deduce all the higher order moments from $\sigma_{x^g}^2 = \mathrm{E}[(x^g)^2]$

$$\begin{aligned} m_{2p+1}^g &= \mathrm{E}[(x^g)^{2p+1}] = 0, \\ m_{2p}^g &= \frac{(2p-1)!}{2^{p-1}(p-1)!}\sigma_{x^g}^{2p}, \end{aligned} \tag{5}$$

where $p > 0$. Similarly, for a zero mean Laplacian process x^l:

$$\begin{aligned} m_{2p+1}^l &= \mathrm{E}[(x^l)^{2p+1}] = 0, \\ m_{2p}^l &= \frac{(2p)!}{2^p}\sigma_{x^l}^{2p}, \end{aligned} \tag{6}$$

where $\sigma_{x^l}^2 = \mathrm{E}[(x^l)^2]$. For a zero mean uniform process x^u:

$$\begin{aligned} m_{2p+1}^u &= \mathrm{E}[(x^u)^{2p+1}] = 0, \\ m_{2p}^u &= \frac{3^p\sigma_{x^u}^{2p}}{2p+1}, \end{aligned} \tag{7}$$

where $\sigma_{x^u}^2 = \mathrm{E}[(x^u)^2]$. As depicted on Fig. 1, the larger is the shape factor ν, the better is the matrix conditioning, for all considered values of N.

Thus, we expect to achieve better nonlinear identification for Gaussian inputs than for Laplacian inputs. Ideally, the uniform distribution provides the best conditioning.

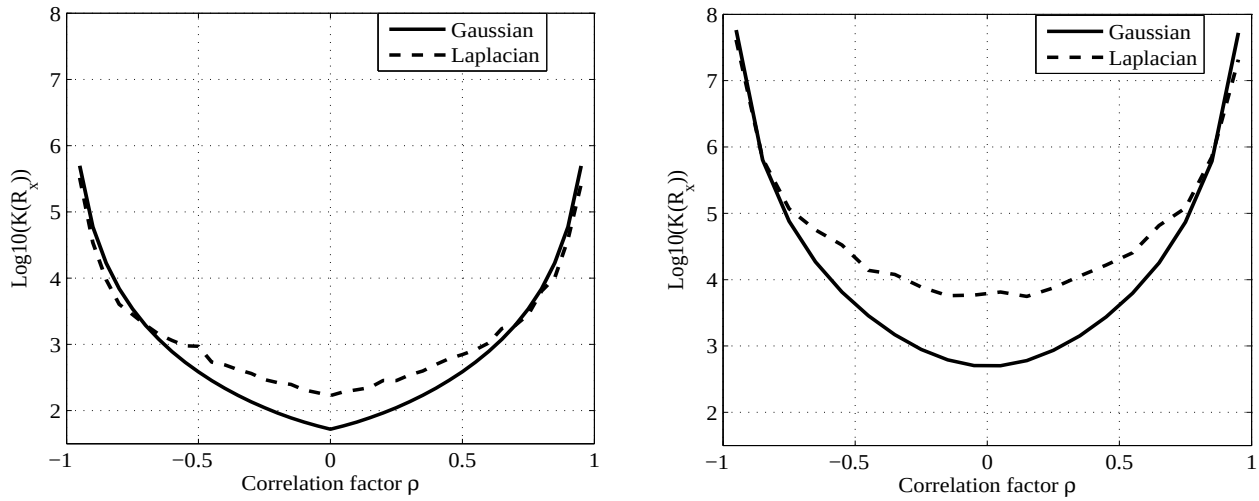

Fig. 2. Condition number of the estimated matrix $\widehat{\mathbf{R}}_{\mathbf{x}}$ (50.000 samples) according to the correlation factor ρ, for Volterra systems S_1 (left, $M = 2$ and $N = 3$) and S_2 (right, $M = 2$ and $N = 2$).

B. Volterra systems

Some nonlinear systems with memory, like loudspeakers, are modeled by Volterra series. Let N be the polynomial order and M the memory length of a Volterra model. We denote by X_k the Volterra input vector defined by

$$X_k = \underbrace{Z_k \oslash Z_k \oslash ... \oslash Z_k}_{N \ \ times}, \tag{8}$$

where $Z_k = [1, x_k, x_{k-1}, ..., x_{k-M+1}]^\top$. $\oslash$ denotes a modified Kronecker product, whose resulting redundant terms are omitted. X_k is a vector of length $\frac{(M+N)!}{N!M!}$ and it contains only products belonging to the set $\{x_k^{m_1} x_{k-1}^{m_2} ... x_{k-M+1}^{m_M} / m_1 + m_2 + ... + m_M \leq N\}$ [21]. As for polynomial systems, we will call the matrix $\mathbf{R}_{\mathbf{x}} = \mathrm{E}[X_k X_k^\top]$ "correlation matrix". It was shown in [22] that for an i.i.d [3] process, the conditioning of the correlation matrix $\mathbf{R}_{\mathbf{x}}$ increases exponentially with the nonlinearity order N and the memory length M and has the upper bound $[K(\mathbf{C}_{\mathbf{x}})]^M$:

$$K(\mathbf{R}_{\mathbf{x}}) < [K(\mathbf{C}_{\mathbf{x}})]^M. \tag{9}$$

Unlike polynomial systems, the observation matrix $\mathbf{R}_{\mathbf{x}}$ for Volterra systems contains auto-correlation terms of the input signal x_k (like $E[x_k x_{k-i}]$) and cross-correlation terms (for example $E[x_k^p x_{k-i}^q]$). For

[3] i.i.d: independent and identically distributed.

correlated processes, an upper bound for the conditioning of the matrix $\mathbf{R_x}$ is difficult to determine theoretically.

To show the influence of the input correlation on the conditioning of the correlation matrix, we compare on Fig. 2 the conditioning of the matrix $\mathbf{R_x}$ for first order correlated Gaussian and Laplacian [23] processes (AR(1)) according to the correlation factor ρ for the two following Volterra systems:

- S_1 : a Volterra system of order $N = 3$ and memory $M = 2$ (containing 15 coefficients)
- S_2 : a Volterra system of order $N = 2$ and memory $M = 2$ (containing 10 coefficients).

As expected, Fig. 2 shows that the condition number increases with the correlation of the input signal in both cases (Gaussian and Laplacian processes). Furthermore, we notice that:

- for low correlation, the correlation matrix is better conditioned for the Gaussian process than for the Laplacian process
- for high correlation, the condition numbers are quietly the same for both processes.

Hence, only for low correlated processes, the input Gaussianity enhances the conditioning of the involved correlation matrices.

In the following, we show that such a property is however required for performance enhancement of nonlinear system identification (polynomial and Volterra).

C. Conditioning enhancement through orthogonalization

A powerful way to improve the conditioning is to orthogonalize the observation matrix $\mathbf{C_x}$ or $\mathbf{R_x}$. For any PDF of the input, this may be achieved through the Gram-Schmidt procedure [9]. If the system is memoryless and the input is Gaussian, the orthogonalization may be performed more simply using a set of Hermite polynomials $\{H_0(x), H_1(x), ..., H_N(x)\}$ [1], where the higher order moments can be expressed using only the signal variance.

If the system has memory (Volterra system) and the input is Gaussian, this holds only if the input is white. In the case of a Gaussian correlated input, the latter has to be whitened as proposed in [17]. The backward prediction errors of respective orders $0, 1, \ldots, M-1$ form the input vector of the new identification system. One can then orthogonalize this vector using Hermite polynomials.

D. Variability of the PDF of audio signals

Audio signals are globally generalized Gaussian but this should be locally verified. We present in Fig. 3 and 4 respectively the PDF of 2000 samples of a speech signal sampled at 8 kHz (250 ms) and a

music signal sampled at 44.1 kHz (45 ms), which vary from one frame to another between Gaussian and Laplacian processes. Particulary, for voiced (speech) or tonal (music) zones, the PDF is near a Gaussian distribution.

Consequently, in an adaptive identification of a nonlinear system, since the performance depends on the local properties of the signal, one may expect this variability of the local PDF to lead to a variability of the conditioning and, consequently, of the identification performance.

E. Conclusion

We have shown in this section that the performance of nonlinear system identification depends on the conditioning of the observation matrix and, hence, on the PDF of the input. For memoryless systems, the flatter is the distribution, the better is the conditioning. Considering generalized Gaussian distributions between Laplacian and Gaussian, as audio can be modeled, this means that the identification should perform better with Gaussian inputs. For systems with memory and correlated input, the conditioning is bad whatever the PDF is. However, the Gausiannity is again a desirable property, since it allows a simple orthonormalization of the input, which minimizes the condition number.

Thus, we propose in the following an audio Gaussianization procedure and an identification scheme based on this built Gaussianity and on an input orthonormalization.

III. AUDIO GAUSSIANIZATION

Since the Gaussianity of the input is a desirable property for nonlinear system identification, we propose in the following a specific "doping" technique to *"force"* audio signals to be Gaussian [24], [25].

A. Gaussianization procedure

The proposed transformation of audio signals from their empirical distribution to a Gaussian distribution is performed over non overlapping frames.

We associate to the sequence X of length L the corresponding empirical cumulative distribution function

$$\begin{aligned} F_X^{emp}(x_k) &= P[X \leq x_k] \\ &= \frac{|\{X \leq x_k\}|}{L}, k = 1, \cdots, L. \end{aligned} \tag{10}$$

The distribution of the signal is turned into the Gaussian distribution with the same mean value m_x and variance σ_x^2 through a histogram equalization similar to the basic one used in image processing [26].

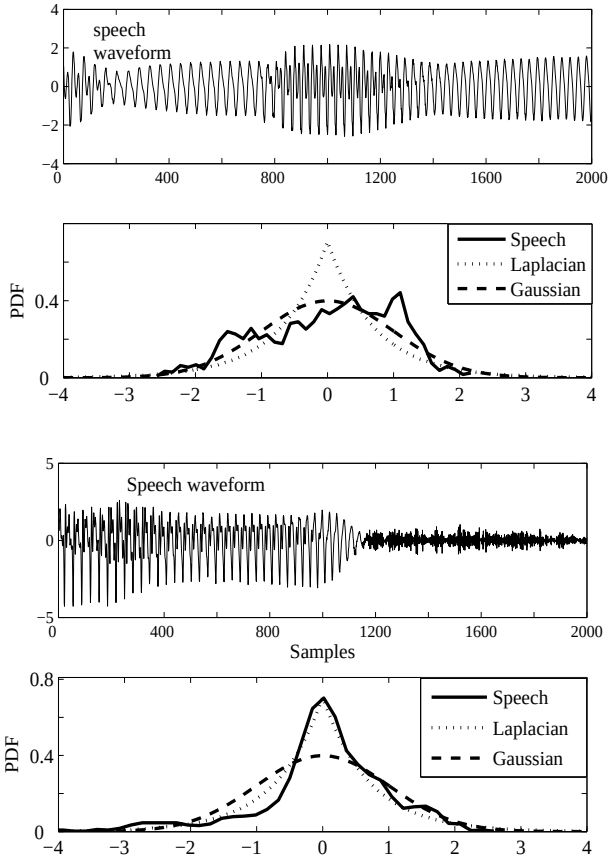

Fig. 3. Speech frames (2000 samples, sampling frequency 8 kHz) with a nearly Gaussian distribution (top) and a nearly Laplacian distribution (bottom).

Denoting F^{target} the cumulative distribution function of the target Gaussian distribution, for $k = 1 \text{ to } L$, we add a small value g_k to each x_k, so that $x_k^w = x_k + g_k$ verifies:

$$F^{target}(x_k^w) = F_X^{emp}(x_k), \tag{11}$$

as shown in Fig. 5. Then we get the Gaussiannized signal

$$x_k^w = x_k + g_k, \tag{12}$$

where g_k is the Gaussianization signal, called the doping watermark [24].

B. Perceptual limits of Gaussianization

To avoid local power peaks of the Gaussianization signal g_k (mainly due to the variability of the short-term PDF), the Gaussianization is performed on long frames, typically $L = 10\ 000$. Thus, the

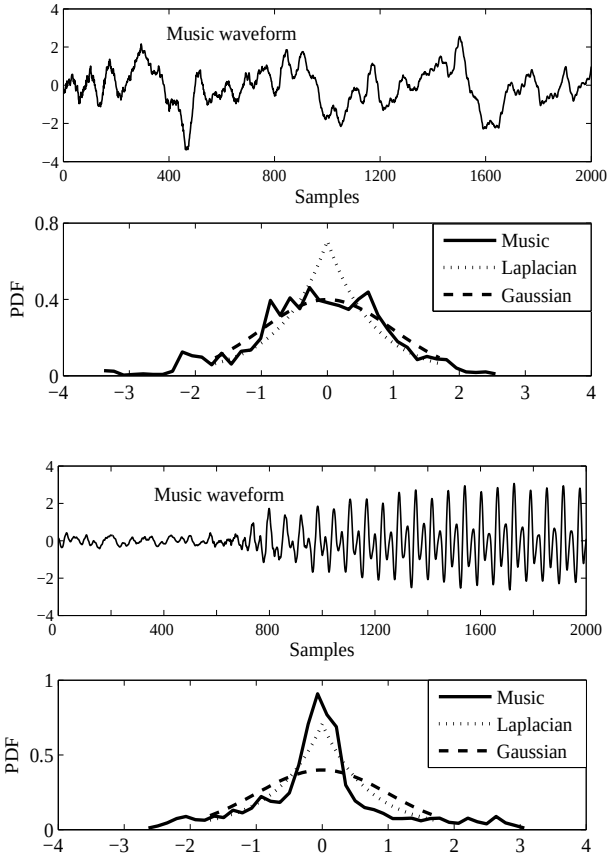

Fig. 4. Music frames (2000 samples, sampling frequency 44.1 kHz) with a nearly Gaussian distribution (top) and a nearly Laplacian distribution (bottom).

Gaussianization is an off-line procedure. However, the inserted Gaussianization signal g_k is clearly audible. One reason is that the PDF of speech and some music signals is much higher than the Gaussian PDF around zero. Consequently, for segments of x with values around zero, the shifts g_k are of the same order as the initial values x_k.

To study the audibility of the Gaussianization signal, we evaluate in Table I the Zero Crossing Rate[4] (ZCR) and the Signal to Gaussianization signal Ratio (SGR) defined as

$$SGR = 10 \log_{10}[P_x/P_g], \tag{13}$$

where P_x and P_g are respectively the power of the signal x_k and the power of the Gaussianization signal g_k. Three different types of segments are then considered: voiced, unvoiced and silent, for speech; tonal,

[4]The zero crossing rate is the ratio between the number of zero crossings and the total number of samples in a signal segment.

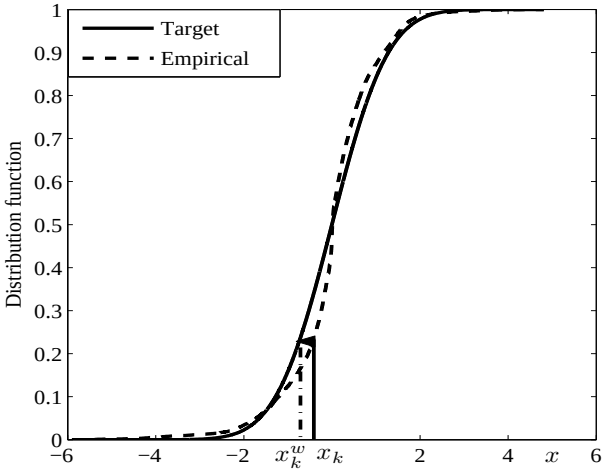

Fig. 5. Gaussian target and empirical cumulative distribution functions for a speech frame of 10 000 samples.

noisy and silent, for music.

As shown on Table I and unsurprisingly, the SGR is the worst for silent segments. The ZCR is much higher for unvoiced segment than for voiced segments, which leads to a worse SGR. Hence, in order to reduce the power of the Gaussianization signal, we proposed to exclude silent and unvoiced/noisy segments from the Gaussianization procedure.

Since this is not sufficient to make the watermark inaudible, we achieved a perceptual masking through an iterative limitation of the variance of the Gaussianization signal.

C. Gaussianization under perceptual constraint

The inaudibility is preserved by reaching a target variance $(\sigma_g^{target})^2$ for the Gaussianization signal g, through an iterative adjustment of the maximum value of $|g|$, denoted g_{max}.

As a first step, we fix an arbitrary authorized maximum value g_{max}, which will be used to define the search interval of a dichotomy process to find the optimal value g_{max}^{opt}. We transform the PDF of x under the constraint $|g| < g_{max}$, which provides a variance σ_g^2 for g.

As a second step, we perform the following test:

- if $\sigma_g < \sigma_g^{target} - \epsilon$ (ϵ is an arbitrary small value) then we fix $g_{max}^1 = g_{max}$ and repeat the multiplication of g_{max} by 2 and the PDF transformation under the constraint $|g| < g_{max}$, until we get $\sigma_g > \sigma_g^{target}$. Let $g_{max}^2 = g_{max}$
- if $\sigma_g > \sigma_g^{target} + \epsilon$ then we fix $g_{max}^2 = g_{max}$ and we repeat the division of g_{max} by 2 and the PDF transformation under the constraint $|g| < g_{max}$ until we get $\sigma_g < \sigma_g^{target}$. Let $g_{max}^1 = g_{max}$.

	ZCR	SGR[dB]
unvoiced segments	0.35	-0.77
silent segments	-	-7.96
voiced segments	0.014	6.46

TABLE I

EVALUATION OF THE ZCR AND THE SGR FOR VOICED, UNVOICED AND SILENT SEGMENTS OF A SPEECH SIGNAL.

Signal	Signal duration (s)	χ_{dB}	Quality
Speech	1.2	-18	estimated MOS=3.87
Pop music	2	-16	ODG=-0.758
Classical music	1.5	-16	ODG=-0.18
Guitar	2.38	-16	ODG=-1.13

TABLE II

PEAQ (ODG) AND PESQ (ESTIMATED MOS) EVALUATIONS FOR SOME AUDIO SIGNALS AFTER GAUSSIANIZATION PROCESSING.

In both cases we get an interval $[g^1_{max}, g^2_{max}]$ in which we search by dichotomy the optimal value g_{max} that verifies $|\sigma_g - \sigma_g^{target}| < \epsilon$.

How to determine σ_g^{target}? Since the PSD of g_k is roughly parallel to the PSD of x_k, the target variance must be at least 13 dB under the variance of x, according to [27]. We set:

$$\sigma_g^{target} = \chi \sigma_x,$$

where the attenuation factor χ was fixed after informal subjective tests and chosen to guarantee the imperceptibility of g.

Finally, after Gaussianization of voiced (for speech) or tonal (for music) segments according to the process described above, we concatenate silent and unvoiced (or non-tonal) segments, which are not Gaussianized.

D. Audio quality evaluation of Gaussianized signals

Perceptually, the Gaussianized signal and the original one must be the same. Audio quality is preferably evaluated through formal subjective measures. Nevertheless, for a rapid and low-cost evaluation, they can

be replaced by objective measures like PEAQ (Perceptual Evaluation of Audio Quality) for music [28] and PESQ (Perceptual Evaluation of Speech Quality) for speech [29].

For PEAQ measures, an ODG (Objective Difference Grade) score is computed which is in $[-4, 0]$. The score 0 indicates an imperceptible difference between the original signal and its processed version. The value -4 refers to the highest degradation level.

For PESQ measures, the quality evaluation is done through an estimated MOS (Mean Opinion Score) which is in $[1, 4.5]$. The value 4.5 corresponds to the best fidelity to the original signal and the value 1 refers to the highest degradation.

The ODG and the estimated MOS values, relative respectively to music and speech signals after Gaussianization under a perceptual constraint fixed through the choice of χ, are displayed in Table II. These results indicate that the Gaussianization modifies slightly the audio quality.

E. Gaussianity measurement

The Gaussianity of a signal may be measured by its Kurtosis, which equals 3 for a Gaussian distribution. For the previous pop-music signal, we estimated the Kurtosis for non-overlapping frames of $10\ 000$ samples, for the original signal x and the signal Gaussianized under the inaudibility constraint expressed by $\chi_{\text{dB}} = -16$ dB. As shown by Fig. 6, the Kurtosis of the Gaussianized signal x^w is closer to 3 than that of the original signal x for most of the frames. **The variability of the estimated Kurtosis around 3 results from the exclusion of the silent and noisy segments from the Gaussianization.**

F. Conclusion

We have shown in section II that the Gaussianity of the input is a desirable property for identification of nonlinear systems. Since audio signals are generally non-Gaussian, we have proposed a Gaussianization method that makes an audio signal more Gaussian (but not fully Gaussian), while ensuring its perceptual fidelity to the original. We show in the following how this leads to higher performance in nonlinear system identification.

IV. Nonlinear audio system identification relying on input Gaussianity

As stressed in [14] for loudspeakers, the nonlinear behavior of audio systems varies in time according to the input signal and to the excited physical effect of the system (heating for example). Hence, the system identification has to be adaptive, in order to observe these variations. Moreover, the transient state of the identification has to be as short as possible in order to observe the early behavior of the system.

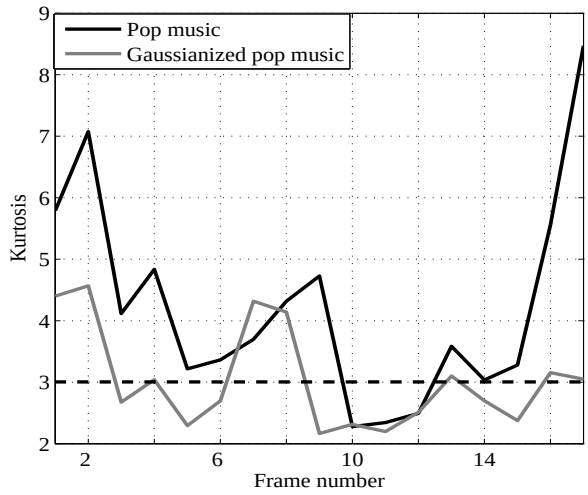

Fig. 6. Estimated Kurtosis evolution of original and Gaussianized signals (non-overlapping frames of 10 000 samples).

However, the proposed Gaussianization is an off-line process, since it is performed on large segments of an audio signal. Thus, for an off-line identification task, the signal has to be fully Gaussianized before identifying the system. For a real-time identification (for example with a compensation purpose), the Gaussianization is suitable in the context of playing/broadcasting recorded material, and not in nonlinear acoustic echo cancellation for example.

A. Classical system identification scheme

We consider here a nonlinear system A which is identified by an adaptive nonlinear filter A_k (polynomial or Volterra model). The input and output of the nonlinear (NL) system are denoted respectively by x_k and y_k and the estimated output $\widehat{y}_k$ is

$$\widehat{y}_k = A_k^\top X_k, \tag{14}$$

where $A_k = [1, a_1, ..., a_q]^\top$. The structure of the input vector X_k and its length q are related to the nonlinear model (for polynomial model $q = N + 1$). The estimation error is

$$e_k = y_k - \widehat{y}_k. \tag{15}$$

A_k is the adaptive filter updated with a normalized Least-Mean Square (NLMS) algorithm [30] and driven by the input vector X_k as follows

$$A_{k+1} = A_k + \frac{\mu}{\|X_k\|^2} e_k X_k, \tag{16}$$

where μ is the adaptation step size.

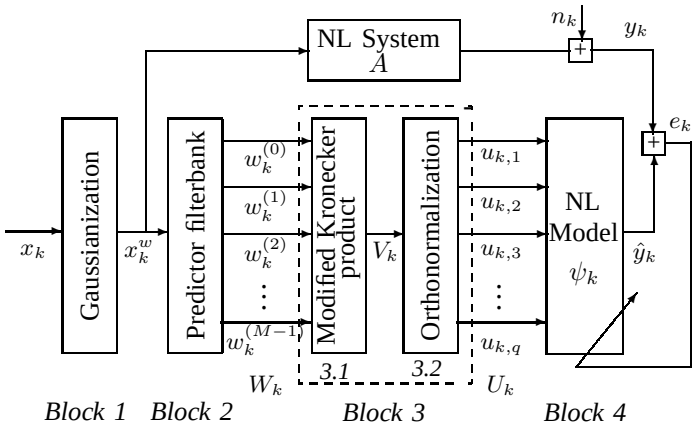

Fig. 7. Proposed identification scheme for NL Volterra systems: Gaussianization, predictor filterbank, orthonormalization and adaptive NL model.

In this paper, the NLMS algorithm was chosen as an example to illustrate the proposed methodology, but other identification algorithms could be used.

For the analysis purpose, we consider here that the system and the model have the same structure. Then, the system output is

$$y_k = A^\top X_k + n_k, \tag{17}$$

where A is the NL system and n_k is an additive white Gaussian noise.

B. The proposed identification structure

Based on the conclusions of sections II and III, we propose the identification method depicted on Fig. 7 for both memoryless NL systems and NL systems with memory. Note that the second block and the first part of the third block concern only the identification of NL systems with memory.

Whereas orthogonalization was already proposed in [17], we propose here to further improve the conditioning of the matrix involved in the identification system (block 4) through an *orthonormalization* step.

1) Gaussianization (block 1): This first block consists in the Gaussianization of the audio signal x_k as detailed above. The input Gaussianity is necessary for the following orthonormalization block.

2) Predictor filterbank (block 2): It consists in computing the prediction errors $\{w_k^{(0)}, w_k^{(1)}, w_k^{(M-1)}\}$ of respective orders $0, 1, ..., M-1$. This step was proposed in [17] to orthogonalize Gaussian correlated

signals for the identification with Volterra structure. Note that this second pre-processing step concerns only systems with memory and it does not need the hypothesis of input Gaussianity.

3) Orthonormalization (block 3): The goal of this block is to form an orthonormal basis fitted to the NL model (polynomial or Volterra) so that the correlation matrix is the identity matrix with an optimal conditioning equal to 1.

- *For memoryless systems*[5], the goal is to form an orthonormal basis relative to the polynomial basis $B^{(N)}(x_k^w) = \{1, x_k^w, (x_k^w)^2, ..., (x_k^w)^N\}$. We first normalize x^w:

$$\tilde{x}_k^w = x_k^w / \hat{\sigma}_{x^w} \tag{18}$$

  where $\hat{\sigma}^2_{x^w}$ is the estimated variance of x^w, computed on quasi-stationary frames (typically 10 to 30 ms for speech). As the input signal x^w is a Gaussianized signal, the corresponding orthogonal polynomial basis is the Hermite polynomial basis $H^{(N)}(\tilde{x}_k^w) = \{H_0(\tilde{x}_k^w), H_1(\tilde{x}_k^w), ..., H_N(\tilde{x}_k^w)\}$, where H_i denotes the i^{th} Hermite polynomial. If the identification is driven by $H^{(N)}(\tilde{x}_k^w)$, its performance depend on the conditioning of $\mathrm{E}[H^{(N)}(\tilde{x}_k^w)H^{(N)}(\tilde{x}_k^w)^\top]$ (diagonal matrix). To get an optimal conditioning, we use the normalized Hermite polynomial basis $\widetilde{H}^{(N)} = \{\widetilde{H}_0, \widetilde{H}_1, ..., \widetilde{H}_N\}$, where:

$$\forall\, i, \quad \widetilde{H}_i(z) = H_i(z)/\sqrt{i!}$$

  In other terms, we form the vector $U_k = \widetilde{H}(\tilde{x}_k^w)$, so that $\mathrm{E}[U_k U_k^\top]$ is the identity matrix, with conditioning equal to 1.

  The relationship between vectors X_k^w and U_k is then

$$U_k = \mathbf{\Gamma} X_k^w, \tag{19}$$

  where $\mathbf{\Gamma}$ is a $(N+1) \times (N+1)$ lower triangular matrix.

- *For systems with memory:* We propose in the following to do some modifications to the previous identification structure for the identification of nonlinear systems with memory, based also on input Gaussianity hypothesis. This idea is inspired from the Wiener G-functionals [2] which are derived from the Volterra kernels by polynomials combination. The statistical orthogonality properties of the involved kernels improve the conditioning of the correlation matrix only for white and Gaussian inputs.

  To overcome the non-Gaussianity and the high correlation of audio signals, we have introduced

[5]Referring to Fig. 7, $V_k = X_k$.

stages of Gaussianization (block 1) and forward prediction filterbank (block 2).

The goal of block 3 is to form an orthonormal basis relative to the vector $W_k = [1, w_k^{(0)}, w_k^{(1)}, w_k^{(M-1)}]^\top$.

We first normalize each $w_k^{(i)}$ as described above (Eq. (18)), which provides $\widetilde{W}_k = [1, \tilde{w}_k^{(0)}, \tilde{w}_k^{(1)}, \ldots, \tilde{w}_k^{(M-1)}]^\top$.

We apply the modified Kronecker product to the vector $\widetilde{W}_k$ to get the vector V_k:

$$V_k = \widetilde{W}_k \oslash ... \oslash \widetilde{W}_k, \tag{20}$$

which elements are of the form $\prod_{i,j}(\tilde{w}_k^{(i)})^j$. From V_k, we derive a new orthonormal vector U_k which elements are of the form $\prod_{i,j} \widetilde{H}_j(\tilde{w}_k^{(i)})$ [2]. As in the memoryless case, $\mathrm{E}[U_k U_k^\top]$ is the identity matrix, with conditioning equal to 1.

Hence, U_k can be written as $U_k = \mathbf{Q}V_k$, where $\mathbf{Q}$ is a lower triangular matrix. Since $\widetilde{W}_k$ may be written as the product of a lower triangular matrix by $Z_k = [1, x_k, x_{k-1}, ..., x_{k-M+1}]^\top$, according to the properties of the modified Kronecker product, V_k is also the product of a lower triangular matrix by X_k (defined by Eq. (8)). Thus, the relationship between vectors X_k and U_k is again $U_k = \mathbf{\Gamma} X_k$, where $\mathbf{\Gamma}$ is a lower triangular matrix. Note that the input Gaussianity hypothesis is necessary only for this orthonormalization step of the vector V_k. This proposed step is less complicated than the proposed method in [17] where the identifcation system is over-parametrized.

4) Adaptive NL model (block 4): The output of the adaptive identification structure of block 4, driven by the vector U_k provided by block 3, is $\widehat{y}_k = \psi_k^\top U_k$. ψ_k is an adaptive filter updated with a NLMS algorithm as follows

$$\begin{aligned} e_k &= y_k - \psi_k^\top U_k \\ \psi_{k+1} &= \psi_k + \frac{\mu^o}{\|U_k\|^2} e_k U_k, \end{aligned} \tag{21}$$

where e_k is the estimation error and μ^o is the step size.

The three previous pre-processing steps give to the new observation vector U_k better orthogonality properties than the initial vector X_k. Indeed, the obtained correlation matrix $E[U_k U_k^\top]$ which drives the identification is theoretically the identity matrix.

C. Studied schemes for comparative performance analysis

Using the classical adaptive identification algorithm NLMS and for exact modeling (same order for the NL system and model), the transient and steady state behaviors can be studied through the time variation of the deviation vector $\Delta A_k = A - A_k$. Using (15), (16) and (17) it is easy to show that

$$\Delta A_{k+1} = \left(\mathbf{I} - \mu \frac{X_k X_k^\top}{\|X_k\|^2} \right) \Delta A_k - \mu n_k \frac{X_k}{\|X_k\|^2}, \tag{22}$$

where $\mathbf{I}$ refers to the identity matrix of rank q.

For the proposed method, we remind that the estimation error is

$$\begin{aligned} e_k &= y_k - \widehat{y}_k \\ &= A^\top X_k^g + n_k - \psi_k^\top U_k. \end{aligned} \tag{23}$$

We denote by A_k^o the adaptive filter that identifies the NL system A in the proposed scheme. A_k^o is updated as

$$A_k^o = \mathbf{\Gamma}_k^\top \psi_k, \tag{24}$$

where $\mathbf{\Gamma}_k$ denotes the transform matrix computed for the signal frame to which belongs the k^{th} sample. Using (21), (23) and (24), we can show that

$$\Delta A_{k+1}^o = \left(\mathbf{I} - \mu^o \frac{\mathbf{\Gamma}_k^\top U_k U_k^\top \mathbf{\Gamma}_k^{-\top}}{\|U_k\|^2} \right) \Delta A_k^o - \mu^o n_k \frac{\mathbf{\Gamma}_k^\top U_k}{\|U_k\|^2}, \tag{25}$$

where $\Delta A_k^o = A - A_k^o$.

We first study the convergence in the stationary case. The NLMS algorithm can be replaced by the LMS algorithm, which means replacing $\mu/\|X_k\|^2$ and $\mu^o/\|U_k\|^2$ by μ and μ^o, respectively, in the previous equations.

In this case, under the independence assumption between X_k and ΔA_k and for a small step size μ, taking the expectation value of both sides of (22) leads to

$$\mathrm{E}[\Delta A_{k+1}] = \left(\mathbf{I} - \mu \mathrm{E}\left[X_k X_k^\top \right] \right) \mathrm{E}[\Delta A_k]. \tag{26}$$

The mean convergence depends on the conditioning of the matrix $\mathrm{E}\left[X_k X_k^\top\right]$ [31].

Similarly, under the independence assumption between U_k and ΔA_k^o and for a small step size μ^o we can deduce from (25) :

$$\mathrm{E}[\Delta A_{k+1}^o] = \left(I - \mu^o \mathrm{E}\left[\mathbf{\Gamma}_k^\top U_k U_k^\top \mathbf{\Gamma}_k^{-\top} \right] \right) \mathrm{E}[\Delta A_k^o]. \tag{27}$$

Since $\mathbf{\Gamma}_k$ is triangular, the conditioning of $\mathrm{E}\left[\mathbf{\Gamma}_k^\top U_k U_k^\top \mathbf{\Gamma}_k^{-\top} \right]$ is the same as that of $\mathrm{E}\left[U_k U_k^\top\right]$, which is equal to 1, so that the proposed method provides the maximal convergence rate.

In the case of natural audio signals, in spite of the orthonormalization step, U_k is not stationary, so that the LMS algorithm is not convenient. Coming back to the NLMS algorithm, equations (26) and (27) become respectively:

$$\mathrm{E}[\Delta A_{k+1}] = \left(I - \mu \mathrm{E}\left[\frac{X_k X_k^\top}{\|X_k\|^2} \right] \right) \mathrm{E}[\Delta A_k]. \tag{28}$$

$$\mathrm{E}[\Delta A^o_{k+1}] = \left(I - \mu^o \mathrm{E}\left[\frac{\mathbf{\Gamma}_k^\top U_k U_k^\top \mathbf{\Gamma}_k^{-\top}}{\|U_k\|^2} \right] \right) \mathrm{E}[\Delta A^o_k]. \tag{29}$$

Hence, the convergence depends on the conditioning of $\widetilde{\mathbf{R}}_{\mathbf{x}} = \mathrm{E}\left[\frac{X_k X_k^\top}{\|X_k\|^2}\right]$ and $\widetilde{\mathbf{R}}_{\mathbf{u}} = \mathrm{E}\left[\frac{U_k U_k^\top}{\|U_k\|^2}\right]$, respectively. One may expect that the latter is better conditioned than the former and thus provides a faster convergence, but this should be verified experimentally.

In the steady state, the identification performance is evaluated through the classical Mean Square Deviation

$$MSD(k) = \mathrm{E}[\|\Delta A_k\|^2].$$

From (22) and under independence hypothesis between X_k and ΔA_k, we get

$$\begin{aligned} \mathrm{E}[\|\Delta A_{k+1}\|^2] &= \mathrm{E}\left[\Delta A_k^\top (I - \mu(2-\mu)\frac{X_k X_k^\top}{\|X_k\|^2}) \Delta A_k \right] + \\ &\underbrace{\mu^2 \sigma_n^2 \mathrm{E}\left[\frac{1}{\|X_k\|^2} \right]}_{P_k^\nu}, \end{aligned} \tag{30}$$

where σ_n^2 denotes the variance of the noise n.
Similarly, from equation (27), we get

$$\begin{aligned} &\mathrm{E}[\|\Delta A^o_{k+1}\|^2] = \\ &\mathrm{E}\left[(\Delta A^o_k)^\top (I - \mu^o(2-\mu^o) \frac{\|\mathbf{\Gamma}_k^{-1} U_k\|^2 \mathbf{\Gamma}_k^{-1} U_k U_k^\top \mathbf{\Gamma}_k^{-\top}}{\|U_k\|^4}) \Delta A^o_k \right] \\ &\underbrace{+(\mu^o)^2 \sigma_n^2 \mathrm{E}\left[\frac{\|\mathbf{\Gamma}_k^\top U_k\|^2}{\|U_k\|^4} \right]}_{P_k^\gamma}. \end{aligned} \tag{31}$$

From equations (31) and (30) one can see that the steady state performances of the proposed method and the classical adaptive filter depend crucially on the instantaneous values of $P_k^\gamma = (\mu^o)^2 \sigma_n^2 \mathrm{E}\left[\|\mathbf{\Gamma}_k^\top U_k\|^2 / \|U_k\|^4\right]$ and $P_k^\nu = \mu^2 \sigma_n^2 \mathrm{E}\left[1/\|X_k\|^2\right]$ respectively. As A_k^o is computed in a more stationary context where $\mathrm{E}[\|U_k\|^2] \simeq 1$, P_k^γ is expected to have smoother variations than P_k^ν for which $\|X_k\|$ presents high and rapid variations.

The experimental protocol presented in table III is used for the following simulations.

V. Simulation results and discussion

A. Memoryless systems

For performance evaluation of the proposed identification structure for polynomial systems, a polynomial system of order $N = 7$ is identified by an adaptive polynomial filter of order N. The system

"**C**"=Classical adaptive identification method with **non Gaussianized** input x_k -identified output $$y_k = A^\top X_k + n_k \quad (32)$$ -block 4 (and block 3.1 if system with memory) - NLMS algorithm $$\begin{aligned} e_k &= y_k - A_k^\top X_k \\ A_{k+1} &= A_k + \frac{\mu}{\|X_k\|^2} e_k X_k. \end{aligned} \quad (33)$$ -Deviation vector in exact modeling $$\Delta A_k = A - A_k \quad (34)$$	"**G**"= classical adaptive identification method with **Gaussianized** input x_k^w -identified output $$y_k = A^\top X_k^w + n_k \quad (35)$$ - blocks 1 and 4 (and block 3.1 if system with memory) - NLMS algorithm $$\begin{aligned} e_k &= y_k - A_k^\top X_k^w \\ A_{k+1} &= A_k + \frac{\mu}{\|X_k^w\|^2} e_k X_k^w. \end{aligned} \quad (36)$$ -Deviation vector in exact modeling $$\Delta A_k = A - A_k \quad (37)$$
"**O**"= proposed adaptive identification method with **original** input x_k (not Gaussianized) -identified output $$y_k = A^\top X_k + n_k \quad (38)$$ - blocks 3.2 and 4 (and blocks 2 and 3.1 if system with memory) - Observation vector: $U_k = \mathbf{\Gamma} X_k$ - NLMS algorithm $$\begin{aligned} e_k &= y_k - A_k^\top U_k \\ \Psi_{k+1} &= \Psi_k + \frac{\mu^o}{\|U_k\|^2} e_k U_k. \end{aligned} \quad (39)$$ -Deviation vector in exact modeling $$\Delta A_k^o = A - A_k^o \quad (40)$$	"**GO**" = proposed adaptive identification method with **Gaussianized** input x_k^w -identified output $$y_k = A^\top X_k^w + n_k \quad (41)$$ - blocks 1, 3.2 and 4 (and blocks 2 and 3.1 if system with memory) - Observation vector: $U_k = \mathbf{\Gamma} X_k^w$ - NLMS algorithm $$\begin{aligned} e_k &= y_k - A_k^\top U_k \\ \Psi_{k+1} &= \Psi_k + \frac{\mu^o}{\|U_k\|^2} e_k U_k. \end{aligned} \quad (42)$$ -Deviation vector in exact modeling $$\Delta A_k^o = A - A_k^o \quad (43)$$

TABLE III

STUDIED SCHEMES FOR PERFORMANCE EVALUATION OF THE PROPOSED IDENTIFICATION STRUCTURE FOR NL SYSTEMS (WITH STEP SIZES μ AND μ^o).

coefficients are generated using a normal distribution with unit variance. The input is a speech signal sampled at 8 kHz and the additive observation noise n_k is white and Gaussian with a variance fixed according to an $SNR = 40$ dB. The Gaussianization is done over non overlapping frames of $10\ 000$ samples under the inaudibility constraint ($\chi = -18$ dB). The variance involved in the computation of the orthonormal basis is estimated on 256 samples frames.

1) Transient behavior analysis: We compare in Fig. 8 the condition numbers of the estimated matrices $\widetilde{\mathbf{R}}_{\mathbf{x}}$ and $\widetilde{\mathbf{R}}_{\mathbf{u}}$ for original and Gaussianized speech computed over frames of 256 samples. The period 256 corresponds to the updating rate of the transformation matrix $\mathbf{\Gamma}$.

As illustrated in Fig. 8, the proposed identification method with a Gaussianized input improves the conditioning[6]. The condition number is reduced by a factor 1000 compared to the classical method with non-Gaussianized speech. The classical method with Gaussianized speech and the orthonormalization without Gaussianization provide intermediate results.

However, even for the Gaussianized speech signal the condition number is not equal to 1. This can be explained by the imperfect Gaussianity of the Gaussianized signal and by the fact that we optimized the conditioning of $\mathrm{E}[U_k U_k^\top]$ and not that of $\widetilde{\mathbf{R}}_{\mathbf{u}} = \mathrm{E}\left[\frac{U_k U_k^\top}{\|U_k\|^2}\right]$.

We display in Fig. 9 the time variations of the MSD related to original and Gaussianized speech with the classical identification method ('C' and 'G' respectively) and with the proposed identification method ('O' and 'GO' respectively) respectively. Fig. 9 shows the enhancement of the convergence rate achieved by the proposed identification structure 'GO'. This is related to the best conditioning of the observation matrix $\widetilde{\mathbf{R}}_{\mathbf{u}}$ for Gaussianized input. Note that the Gaussianity does not seem to be as crucial as the identification structure for the convergence rate, though this structure relies on a Gaussian hypothesis.

2) Steady state analysis and performances: The steady state performances are also studied here through the MSD, but after convergence.

From equations (31) and (30), the steady state performances of the proposed method and of the classical adaptive filter depend crucially on the instantaneous values of $P_k^\gamma = (\mu^o)^2\sigma_n^2\mathrm{E}\left[\|\mathbf{\Gamma}_k^\top U_k\|^2/\|U_k\|^4\right]$ and $P_k^\nu = \mu^2\sigma_n^2\mathrm{E}\left[1/\|X_k\|^2\right]$ respectively. These quantities are displayed in Fig. 10.

As expected, P_k^γ has smoother variations and lower values than P_k^ν. Consequently, as illustrated in Fig. 9, the MSD in the steady state reaches lower values with the proposed method.

[6]Note that on Fig. 8 the conditioning peak at frame 7 corresponds to silent zones where the Gaussianization has no effect.

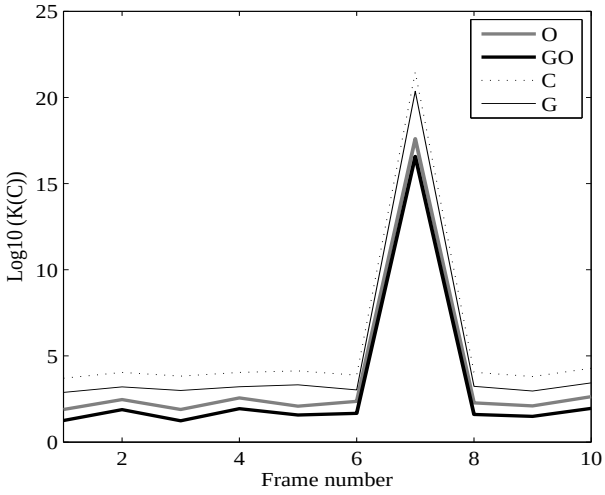

Fig. 8. For a polynomial system, condition number of the estimated matrices $\widetilde{\mathbf{R}}_{\mathbf{x}}$ (schemes 'C' and 'G') and $\widetilde{\mathbf{R}}_{\mathbf{u}}$ (schemes 'O' and 'GO'), computed over 32ms frames.

B. Systems with memory

A Volterra system of order $N = 3$ and memory $M = 3$ is considered in the following under the same simulation conditions as in subsection V-A. Knowing that speech signals are highly correlated and non stationary, the prediction errors are computed over 20 ms frames where speech is assumed locally stationary. The same updating rate is imposed to the transformation matrix $\mathbf{\Gamma}$.

1) Transient behavior analysis: First, the Volterra system is identified by an adaptive Volterra filter with $N = 3$ and $M = 3$ (same order and memory). To point out the enhancement of the convergence rate achieved by the proposed identification structure, we plot on Fig. 11 the time variations of the MSD for the proposed identification structure and the classical adaptive identification where the system is excited by speech signal without pre-processing ('C' or 'O') or by Gaussianized speech ('G' or 'GO').

Fig. 11 shows that the best convergence rate is obtained for the proposed identification structure driven by the Gaussianized speech signal. The enhancement achieved by the proposed identification structure in the transient state is due to the better conditioning of the matrix $\widetilde{\mathbf{R}}_{\mathbf{u}}$ compared to the matrix $\widetilde{\mathbf{R}}_{\mathbf{x}}$ as shown in Fig. 12. Note that the compliance with the Gaussian hypothesis is more crucial here than in the memoryless example.

However, the conditioning of the estimated matrix for Gaussianized speech with the proposed identification structure is not equal to 1, for the reasons given in the case of a memoryless system and because the prediction errors are not perfectly orthogonal.

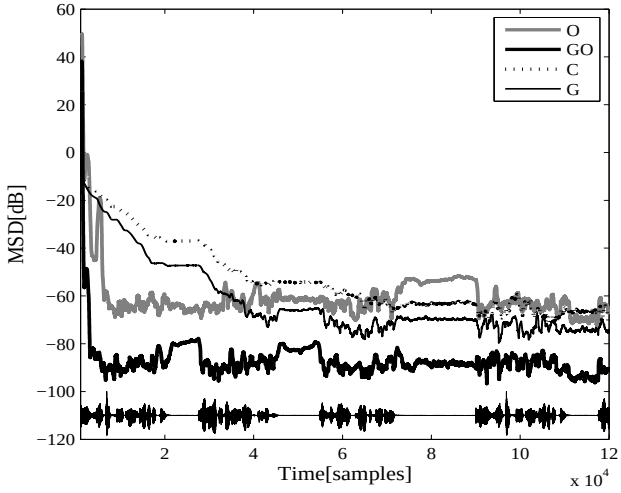

Fig. 9. For a polynomial system, MSD time variation for classical identification and for the proposed method, for original speech and Gaussianized speech. ($N = 7$, $\mu = \mu^o = 0.02$, $SNR = 40$ dB and $\chi = -18$ dB).

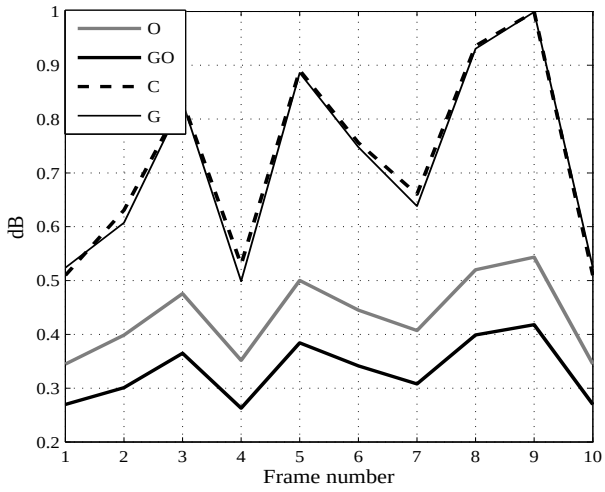

Fig. 10. For a polynomial system,time variation of P^{γ} ('O' and 'GO') and P^{ν} ('C' and 'G') for 256 samples frames ($N = 7$, $\mu = \mu^o = 0.02$, $SNR = 40$ dB and $\chi = -18$ dB).

2) Steady state performances: We display in Fig. 13 the time evolution of P^{γ} and P^{ν}. The same analysis as in the previous case stands. Thus, the proposed identification scheme provides the lowest MSD in the steady state.

Loudspeakers are modeled as nonlinear systems with memories longer than 3 [4], [32]. For a sampling frequency of 44.1 kHz, a memory length of 256 (*ca.* 6 ms) was used in [32]. To point out the effectiveness

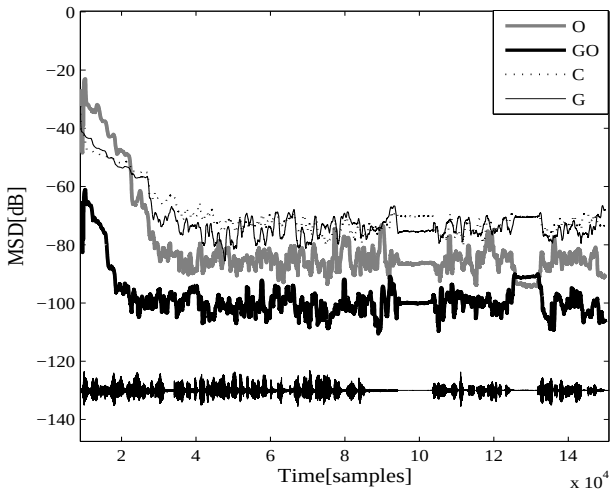

Fig. 11. For a system with memory, MSD time variation for classical identification and for the proposed method, for original speech and Gaussianized speech ($\mu = 0.1$, $N = 3$, $M = 3$, $SNR = 40$ dB and $\chi = -18$ dB).

of the proposed identification scheme in such a more realistic case, we identified a Volterra system of nonlinearity order $N = 3$ and memory length $M = 50$ (6 ms for 8 kHz sampling frequency). The performance is evaluated through the SER [7] (Signal to Error Ratio) measure.

Fig. 14 displays the SER time evolution in steady state (after convergence) of the four studied identification schemes of table III. The enhancement of the proposed identification scheme is ensured even for this larger memory system, where a gain of *ca.* 15 dB is reached most of the time compared to the classical identification without Gaussianization.

3) Under-modeling case: For a more realistic situation of under-modeling, a Volterra system ($N = 3$ and $M = 50$) is identified by a Volterra filter of order $N = 2$ and memory $M = 40$. Hence, we identify only 903 coefficients from all of the 24804 system coefficients. We compare in Fig. 15 the SER time evolution of the four studied identification schemes. A gain of *ca.* 6 dB is reached most of the time, compared to the classical identification without Gaussianization, and *ca.* 2 dB compared to orthonormalization without Gaussianization. Then, the proposed adaptive identification structure guarantees a noticeable enhancement of the identification quality in both cases of exact modeling and under-modeling.

[7]The SER is defined as $SER = 10\log(P_x/P_e)$ where $P_x = \mathrm{E}[x_k^2]$ is the signal power and $P_e = \mathrm{E}[e_k^2]$ is the estimation error power.

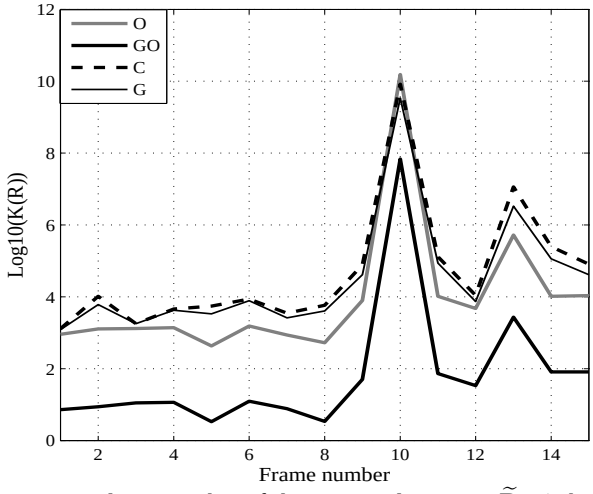

Fig. 12. For a system with memory, condition number of the estimated matrices $\widetilde{\mathbf{R}}_{\mathbf{x}}$ (schemes 'C' and 'G') and $\widetilde{\mathbf{R}}_{\mathbf{u}}$ (schemes 'O' and 'GO'), computed over 32ms frames.

VI. Conclusion

Nonlinear audio system identification methods generally do not take into account audio characteristics: non-stationarity, non-Gaussianity and high correlation.

We have proposed an identification structure suitable for memoryless systems (of polynomial type) and systems with memory (of Volterra type) fitted to these audio properties.

The proposed identification scheme combines audio Gaussianization, whitening and orthonormalization relying on the Gaussianity. We have shown that this pre-processing of the input of an adaptive filter enhances significantly the convergence rate and the identification performance in steady state.

Because of the inaudibility constraint of the Gaussianization, the signal after this step does not fully match the Gaussianity hypothesis assumed by the following steps of the process, which reduces the identification performance, compared to a perfectly Gaussian signal. This constraint however stands only if the NL system must be identified in real-time, for example for a NL-compensation purpose. In the case of an off-line identification (eg. loudspeaker characterization), the noise added by the Gaussianization does not need to be inaudible, which allows a perfect Gaussianity. One should however verify that the amount of added noise does not significantly change the physical behavior of the NL system, compared to the original signal.

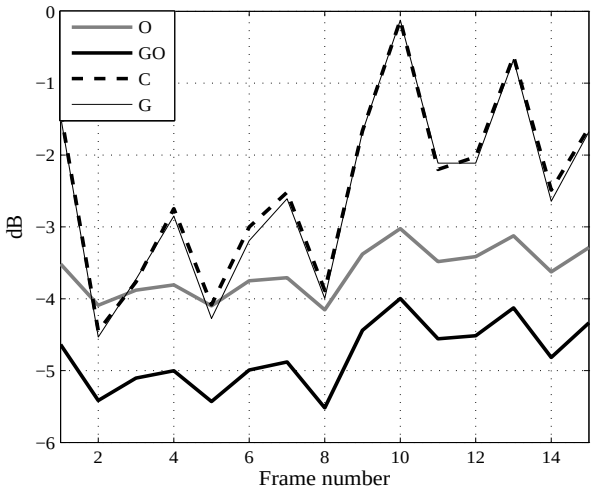

Fig. 13. For a system with memory, time variation of P^{γ} (O and GO) and P^{ν} (C and G) for original and Gaussianized speech for 256 samples frames.

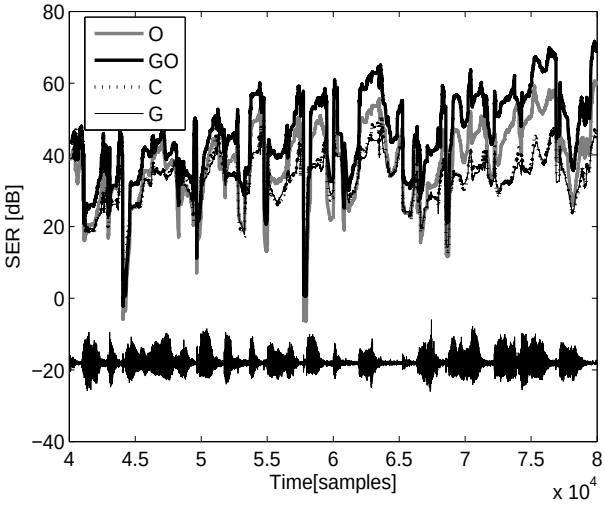

Fig. 14. NL system with memory ($N = 3$, $M = 50$), exact modeling case: Signal to Error Ratio for the classical and the proposed identification methods, for speech with and without Gaussianization ($\mu = 0.1$, $SNR = 40$ dB and $\chi = -18$ dB).

APPENDIX
MODIFIED KRONECKER PRODUCT

A. Volterra models

The principle of Volterra structures is to represent any nonlinear, causal and time invariant system with finite memory by finite Volterra series [2]. For a system with memory M, we consider the following

truncated model of order N

$$y_k = \sum_{j=1}^{N} \left[\sum_{i_1=0}^{M-1} ... \sum_{i_j=0}^{M-1} h_j(i_1, ..., i_k).x_{k-i_1}...x_{k-i_j} \right], \tag{44}$$

where x_k, y_k and h_j represent respectively the system input, its output and the Volterra kernel of order j. Note that in equation (44), there are redundant terms of the form $x_{k-i_1}...x_{k-i_j}$.

B. Mathematical representation of a Volterra filter

The input-output relationship (44) is equivalent to

$$y_k = \Theta^\top X_k, \tag{45}$$

where Θ is the vector containing unique coefficients (after merging redundant terms) of kernels and X_k contains the corresponding products of the input signal necessary for output evaluation. It can be represented through the input vector corresponding to the linear part

$$Z_k = [1, x_k, x_{k-1}, ..., x_{k-M+1}]^\top \tag{46}$$

as

$$X_k = \Omega(\underbrace{Z_k \otimes Z_k \otimes ... \otimes Z_k}_{N terms}), \tag{47}$$

where $\otimes$ denotes the Kronecker product and Ω is the transformation eliminating the redundant terms.

The modified Kronecker product of the n-dimensional vector $Y = [y_1, ..., y_n]^\top$, denoted by $Y \oslash Y$, is the sub-vector of $Y \otimes Y$ of dimension $\frac{n(n+1)}{2}$ as

$$Y \oslash Y = \Omega(Y \otimes Y). \tag{48}$$

This vector representation is used in this article.

REFERENCES

[1] M. Schetzen, *The Volterra and Wiener theories of nonlinear systems*, John Wiley and Sons, 1980.

[2] W. J. Rugh, *Nonlinear system theory: The Volterra/Wiener approach*, Johns Hopkins University Press, 1991.

[3] V. J. Mathews and G. L. Sicuranza, *Polynomial Signal Processing*, John Wiley and Sons, 2001.

[4] V. J. Mathews, "Equalization and linearization of nonlinear systems," in *ICASSP*, 1998.

[5] A. J. M. Kaizer, "Modeling of the nonlinear response of an electrodynamic loudspeaker by a Volterra series expansion," *J. Audio Eng. Soc*, vol. 35, no. 6, pp. 421–433, 1987.

[6] W. Klippel, "Modeling the nonlinearities in horn loudspeakers," *J. Audio Eng. Soc*, vol. 44, no. 6, pp. 470–480, 1996.

[7] L. Tronchin, "The emulation of nonlinear time-invariant audio systems with memory by means of Volterra series," *J. Audio Eng. Soc*, vol. 60, no. 12, pp. 984–996, 2012.

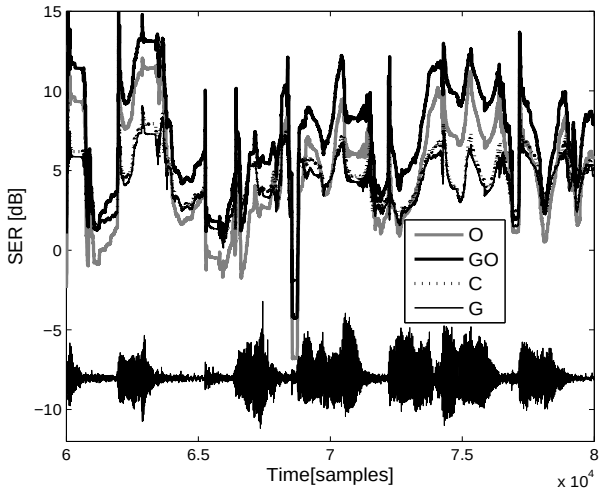

Fig. 15. NL system with memory (N=3, M=50), under-modeling case ($N = 2$, $M = 40$): Signal to Error Ratio for the classical and the proposed identification methods, for speech with and without Gaussianization ($\mu = 0.1$, $SNR = 40$ dB and $\chi = -18$ dB.)

[8] A. Stenger and W. Kellermann, "Adaptation of a memoryless preprocessor for nonlinear acoustic echo cancelling," *Signal Process., Elsevier*, vol. 80, no. 9, pp. 1747–1760, March 2000.

[9] B. Ninness and F. Gustafsson, "A unifying construction of orthonormal bases for system identification," *IEEE Trans. Autom. Control*, vol. 42, no. 4, pp. 515–521, Dec. 1997.

[10] W. Klippel, "Nonlinear system identification for horn loudspeakers," *J. Audio Eng. Soc*, vol. 44, no. 10, pp. 811–820, 1996.

[11] J. Kemp and H. Primack, "Impulse response measurement of nonlinear systems: Properties of existing techniques and wide noise sequences," *J. Audio Eng. Soc*, vol. 59, no. 12, pp. 953–963, 2011.

[12] M. P. Hollier, M. J. Hawksford, and D. R. Guard, "Characterization of communications systems using a speechlike test stimulus," *J. Audio Eng. Soc*, vol. 41, no. 12, pp. 1008–1021, 1993.

[13] W. Klippel, "Loudspeaker nonlinearities- causes, parameters, symptoms," *J. Audio Eng. Soc.*, vol. 54, no. 10, pp. 901–939, 2006.

[14] W. Klippel, "Hot and nonlinear − loudspeakers at high amplitudes, tutorial," in *131st AES Convention*, 2011.

[15] R. van der Zee and A. J. M. Van Tuijl, "Test signals for measuring the efficiency of audio amplifiers," in *104th AES Convention*, 1998.

[16] F. Kuech, M. Zeller, and W. Kellermann, "Input signal decorrelation applied to adaptive second-order Volterra filters in the time domain," in *IEEE Digital Signal Process. Workshop*, 2006.

[17] V. J. Mathews, "Orthogonalization of correlated Gaussian signals for Volterra system identification," *IEEE Signal Process. Lett.*, vol. 2, no. 10, pp. 188–190, Oct. 1995.

[18] S. Gazor and W. Zhang, "Speech probability distribution," *IEEE Signal Process. Lett.*, vol. 10, no. 7, pp. 204–207, July 2003.

[19] G. H. Golub and C. F. Van Loan, *Matrix Computations*, Johns Hopkins University Press, 3rd edition, 1996.

[20] A. Papoulis and S. Pillai, *Probability, random variables and stochastic processes*, Mc Graw Hill, 4th edition, 2002.

[21] T. Ogunfunmi and S. L. Chang, "Second-order adaptive Volterra system identification based on discrete nonlinear Wiener model," *IEE Proc. Vision, Image and Signal Process.*, vol. 148, no. 1, pp. 21–29, Feb. 2001.

[22] R. D. Nowak and B. D. Van Veen, "Random and pseudorandom inputs for Volterra filter identification," *IEEE Trans. Signal Process.*, vol. 42, no. 8, pp. 2124–2135, Aug. 1994.

[23] W. J. Szajnowski, "Generation of a discrete-time correlated Laplacian process," *IEEE Signal Process. Lett.*, vol. 7, no. 3, pp. 69–70, March 2000.

[24] I. Marrakchi, G. Mahé, M. Jaïdane, S. Larbi, and M. Turki, "Gaussianization method for identification of memoryless nonlinear audio systems," in *Proc. European Signal Process. Conf.*, Poland, 2007, pp. 2316–2320.

[25] S. Djaziri Larbi, G. Mahé, I. Marrakchi, M. Turki, and M. Jaïdane, "Doping and witness watermarking for audio processing," in *In Proc. IEEE 7th International Workshop on Systems, Signal Process. and their Applications (WoSSPA)*, Tipaza, Algeria, 2011.

[26] B. Aiazzi, L. Alparone, and S. Baronti, "Estimation based on entropy matching for generalized Gaussian PDF modeling," *IEEE Signal Process. Lett.*, vol. 6, no. 6, pp. 138–140, June 1999.

[27] B. Paillard, P. Mabilleau, S. Morisette, and J. Soumagne, "Perceval: perceptual evaluation of the quality of audio signals," *J. Audio Eng. Soc.*, vol. 40, no. 12, pp. 21–31, Feb. 1992.

[28] "Method for objective measurement of perceived audio quality," ITU-R Rec. BS.1387-1, 2001.

[29] "Perceptual evaluation of speech quality PESQ, an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," ITU-T Rec. P.862, 2001.

[30] N. Bershad, "Analysis of the normalized LMS algorithm with Gaussian inputs," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 34, no. 4, pp. 793–806, 1986.

[31] S. Haykin, *Adaptive Filter Theory*, Prentice-Hall, 1991.

[32] M. Tsujikawa, T. Shiozaki, Y. Kajikawa, and Y. Nomura, "Identification and elimination of second-order nonlinear distortion of loudspeaker systems using Volterra filter," in *IEEE ISCAS*, Geneva, 2000.

Imen Mezghani-Marrakchi received the engineering degree in electrical engineering in 2002 and the M.Eng. degree in 2004 both from the Ecole Nationale d'Ingénieurs de Tunis (ENIT), Université Tunis El Manar, Tunisia. She obtained the PhD degree in telecommunications from the Université Paris Descartes, France, and from ENIT in 2010. She is currently assistant Professor at Ecole Nationale dIngénieurs of Sousse, Tunisia. She is a researcher at the Signals and Systems Lab at ENIT. Her teaching and research interests are in signal processing, audio processing and audio watermarking for linear and nonlinear audio system identification.

Gaël Mahé received the engineering degree in telecommunications and the M.Sc. in signal and communications from Télécom Bretagne, France, in 1998. From 1999 to 2002 he has worked at Orange Labs in Lannion, France, where he received the PhD in signal and telecommunications from the University of Rennes 1 in 2002. Since 2003, he has been with the Laboratory of Informatics of Paris Descartes University (LIPADE), where he is currently assistant Professor. His research deals mainly with new uses of watermarking in audio processing.

Sonia Djaziri-Larbi received the Dipl. Ing. degree in electrical engineering from the Friedrich Alexander Universität of Erlangen-Nürnberg, Germany, in 1996 and the M.Eng. from the Ecole Nationale d'Ingénieurs de Tunis (ENIT), Tunisia, in 1999. She obtained the PhD degree in telecommunications from the Ecole Nationale Supérieure des Télécommunications of Paris and from ENIT in 2005. Since 2001, she has been with the Information and Communications Technologies Department at ENIT, where she is currently assistant Professor. She is a researcher at the Signals and Systems Lab at ENIT, Université Tunis El Manar. Her teaching and research interests are in signal processing, audio processing and audio watermarking.

Mériem Jaïdane received the M.Sc. degree in electrical engineering from Ecole Nationale d'Ingénieurs de Tunis (ENIT), Tunisia, in 1980. From 1980 to 1987, she has worked as research engineer at the Laboratoire des Signaux et Systèmes, CNRS/Ecole Supérieure d'Electricité, France. She received the Doctorat d'Etat degree in 1987. Since 1987, she has been with ENIT where she is currently a full Professor at the Information and Communications Technologies Department. She is a researcher at the Signals and Systems Lab at ENIT, Université Tunis El Manar. Her teaching and research interests are in adaptive systems for digital communications and audio processing.

Monia Turki-Hadj Alouane was born in Mahdia, Tunisia. She received the Principal Eng., M. Eng. and the Ph.D. degrees all from the Department of Electrical Engineering, the Ecole Nationale d'Ingénieurs de Tunis (ENIT), Tunisia, in 1989, 1991 and 1997 respectively. She joined the Department of Information and Communications Technologies of ENIT since 2001, where she is currently full Professor. Since 2010, she is the head of the Signals and Systems Lab at ENIT, Université Tunis El Manar.